On-the-fly Finger-Vein-based Biometric Recognition using Deep Neural Networks

Rıdvan Salih Kuzu, *Student Member, IEEE*, Emanuela Piciucco, *Student Member, IEEE*, Emanuele Maiorana, *Senior Member, IEEE*, and Patrizio Campisi, *Senior Member, IEEE*

***Abstract*—Finger-vein-based biometric recognition technology has recently attracted the attention of both academia and industry because of its robustness against presentation attacks and the convenience of the acquisition process. As a matter of fact, some contactless vein-based recognition systems have already been deployed and commercialized. However, they require the users to keep their hands still over the acquisition device for a few seconds to perform recognition. In this study, we release this constraint and allow users to have their finger vein patterns acquired on the fly. To accomplish this goal, we introduce an ad-hoc acquisition architecture capable of capturing the finger vein structure using an array of low-cost cameras, and we propose a recognition framework based on the use of convolutional and recurrent neural networks. To test the proposed approach we acquire a finger vein image dataset, in video format at four different exposure times, from 100 subjects. The obtained experimental results show that, even in a very challenging scenario, the proposed system guarantees high performance levels, up to 99.13% recognition accuracy over the collected dataset.**

***Index Terms*—Finger Vein Biometrics, Multimodal Biometrics, Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory Networks.**

I. Introduction

In the last decades, the use of biometric technology has emerged as a prominent and well-accepted solution in secure and user-convenient recognition applications [1]. Biometric characteristics such as fingerprint, face, signature, iris, and voice have been successfully employed for access control, border control, financial-related applications, and verification on personal devices, to cite a few examples. However, many of the mentioned biometric identifiers are exposed to the public and therefore prone to presentation attacks [2]. On the contrary, vein patterns are subcutaneous structures and thus intrinsically more robust to such threats than most of the commonly employed biometric traits. Therefore, in the last few years, the use of vein patterns has become more and more popular, and commercial devices have been deployed for real-life applications.

A standard device for the acquisition of vein patterns is composed of a near-infrared (NIR) illuminator and a NIR camera. The haemoglobin in the blood absorbs infrared light, thus revealing the structure of the vessels as dark lines. In our work, we take a step forward and introduce, for the first time in literature, an innovative on-the-fly, contactless, and low-cost vein-based biometric recognition system. The acquisition of the finger vein structure is carried out while the user is on the move, with an increase of user convenience and recognition system throughput. Deep learning approaches, based on both convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are here exploited to extract discriminative features from the acquired vein patterns, and achieve remarkable recognition performance.

The paper is structured as follows: an introduction on finger-vein-based recognition is given in Section II. An overview of the state-of-the-art works using deep-neural-network (DNN) approaches in the framework of vein-based biometric recognition is provided in Section III. Section IV describes the overall identification pipeline of the proposed system, detailing its hardware and software components, while the adopted DNN-based recognition process is outlined in Section V. Finally, the achieved results and conclusions are presented in Sections VI and VII, respectively.

II. Finger-Vein-based Biometric Recognition

The convenience of the data acquisition process, the higher security in terms of presentation attack and liveness detection with respect to conventional biometric traits, as well as the always-improving recognition performance, are leading to an increasing interest in vein-based recognition systems. In the recent years, finger vein [3], palm vein [4], [5], wrist vein [6], and hand-dorsal vein [7], [8], have been studied in the context of biometric recognition [9]. The approaches proposed in the literature capture the vein-patterns either requiring the subject to touch a support, as for Hitachi's VeinID finger vein technology, or allowing a contactless acquisition as in Fujitsu's PalmSecure^{TM}. This technology has already been deployed in real-life applications such as banking or consumer products. However, it is worth pointing out that all the existing solutions require the users to hold their hands still during the entire acquisition process. On the contrary, in our approach, as detailed in the following sections, we allow the users to pass their hands over the sensor while walking, thus implementing a contactless and on-the-fly interaction modality.

Roughly speaking, a vein-based biometric recognition system includes the data acquisition, preprocessing, feature extraction, and classification modules. The vein traits are cap-

The authors are with the Section of Applied Electronics, Department of Engineering, Roma Tre University, 00146 Rome, Italy. (e-mail: {ridvansalih.kuzu, emanuela.piciucco, emanuele.maiorana, patrizio.campisi}@uniroma3.it).

Please cite this work as: Ridvan Salih Kuzu, Emanuela Piciucco, Emanuele Maiorana, Patrizio Campisi, "On-the-Fly Finger-Vein-Based Biometric Recognition Using Deep Neural Networks," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 2641 - 2654, February 2020.

Digital Object Identifier 10.1109/TIFS.2020.2971144

TABLE I
STATE-OF-THE-ART WORKS ABOUT FINGER-VEIN-BASED BIOMETRIC IDENTIFICATION SYSTEMS.

Paper	Database			Employed System		Performance
	Name	# Classes	Categ.	Feature Extraction	Matching	
Kumar et al. [10]	HKPU [10]	302 (156 users)	GB	Gabor filter & morphological proc.	XOR-based similarity scores	CIR = 90.08%
Van et al. [11]	SDUMLA [12]	636 (106 users)	SL	MFRAT [13] & GridPCA	Euclidian distance	CIR = 95.67%
Lu et al. [14]	SDUMLA [12]	636 (106 users)	SB	Polydirectional LLBP	Histogram intersection	CIR = 99.21%
Ong et al. [15]	SDUMLA [12]	636 (106 users)	GB	Minutiae	Genetic algorithm & k-modified Hausdorff distance (k-MHD)	CIR = 99.70%
Qui et al. [16]	SDUMLA [12]	636 (106 users)	SL	Pseudo-elliptical transform & 2D-PCA	Euclidean distance	CIR = 97.61%
	FV-USM [17]	492(123 users)				CIR = 97.02%
Xie et al. [18]	SDUMLA [12]	636 (106 users)		Block-based average absolute deviation (AAD)	Ensemble component-based extreme learning machines (EC-ELM) network	CIR = 97.76%
Banerjee et al. [19]	SDUMLA [12]	636 (106 users)	GB	Morphologically enhanced images	Affine registration based template matching algorithm (ARTeM)	CIR = 90.72%
Yang et al. [20]	HKPU [10]	302 (156 users)	GB	Anatomy Structure Analysis based Vein Extraction(ASAVE)	Integration Matching	CIR = 99.68%
Yang et al. [21]	HKPU [10]	302 (156 users)	GB	ASAVE [20] & indexing	Grouped Hamming Distance	CIR = 97.89%
	SDUMLA [12]	636 (106 users)				CIR = 95.25%
	MMCBNU_6000 [22]	600(100 users)				CIR = 94.83%
	FV-USM [17]	492(123 users)				CIR = 98.31%

tured by using a NIR illuminator and a NIR camera. Rotation and vertical translation of the hand could affect the acquisition step, and the acquired pictures are generally characterized by low contrast and poor quality. This implies the need of some preprocessing steps, namely, registration, normalization and image enhancement. In addition, a region of interest (ROI), containing the vein pattern, is usually extracted from the enhanced image and analyzed in the feature extraction stage. The feature extraction approaches used in vein-based biometrics can be broadly categorized into five classes:

- **geometry-based methods** (GB): geometric information, such as shape or topological structure, is extracted from the vein images and used as discriminative features. Most of such techniques segment vein patterns from the background and then extract relevant features from the obtained network. Global topology-based methods, such as line-like [23], [24] and curvature [25] models, as well as local geometry-based models, such as vein knuckle shapes, endpoints and crossing points of vein structures [26]–[28], belong to this category;
- **subspace-learning-based methods** (SL): techniques exploiting appearance-based methods, such as linear discriminant analysis (LDA) [29], principal component analysis (PCA) [30], [31] or two-dimensional PCA [32], [33] can be listed within this class;
- **statistical-based techniques** (SB): statistical features as the local binary histogram and moments are exploited to extract discriminative information from vein structures. Methods based on local binary patterns (LBPs) [34], [35], local derivative patterns (LDPs) [36], and invariant moments [37] are examples of feature extraction techniques based on local statistics;
- **local invariant-feature-based methods** (LI): methods inspired by approaches stemming from computer vision, such as those using key points for the scale invariant feature transform (SIFT) as features [38], [39] belong to this category;
- **DNN-based models** (DM): a DNN consists of a sequence of processing layers and can be exploited as feature extractor or classifier module in a vein-based biometric system. Since DNNs are used in our approach to process the acquired vein patterns, a detailed overview of the state of the art exploiting DNN-based models for vein-based biometric recognition system is provided in Section III.

An overview of the state of the art on finger-vein-based biometric identification systems is provided in Table I, where details about the employed databases, feature extraction techniques, feature categories, and matching algorithms, together with the achieved recognition performance expressed as correct identification rate (CIR), are reported.

III. STATE OF THE ART: DNN- AND VEIN-PATTERN-BASED BIOMETRIC APPLICATIONS

In the recent past, there has been an increase in the use of deep learning techniques for biometric recognition purposes [60], due to the good performance they achieve. In this section, we present an overview of the most relevant papers using deep learning methods in the field of vein-based biometrics. The related details are summarized in Table II.

A deep learning approach applied to a finger-vein-based biometric identification system has been first proposed by Radzi *et al.* [40]. The structure of the employed network relies on the one presented in [41], with the CNN fed with binary images obtained by thresholding the original vein images. A more recent work on finger-vein-based identification using CNNs is the one proposed by Das *et al.* [3], where stable and highly-accurate performance is achieved while dealing with finger vein images of different quality. Hong *et al.* [43] have designed a finger-vein-based verification system exploiting a pre-trained model of VGG-16 [44]. The pre-trained network is used for fine-tuning, having the difference between two finger vein images as input. Databases with different image qualities are taken into account. A deep CNN (D-CNN) architecture, inspired by VGG-16, has also been exploited for finger-vein-based biometric verification by Huang *et al.* [45]. The modified CNN architecture is fed with a two-channel image resulting

TABLE II
STATE-OF-THE-ART WORKS ABOUT APPLICATIONS OF DEEP LEARNING ALGORITHMS IN THE FIELD OF VEIN-BASED BIOMETRIC RECOGNITION SYSTEMS.

Paper	Biometric Identifier	Database Name	Database # Classes	CNN Features Reference	CNN Features Modality	Performance
Radzi et al. [40]	Finger vein	Own	300 (50 users)	[41]	Biometric Identification	CIR = 100%
Das et al. [3]	Finger vein	HKPU [10]	302 (156 users)	-	Biometric Identification	CIR = 95.32%
		FV-USM [17]	492 (123 users)			CIR = 97.53%
		SDUMLA [12]	636 (106 users)			CIR = 97.48%
		UTFVP [42]	360 (60 users)			CIR = 98.33%
Hong et al. [43]	Finger vein	Own (Good Quality)	120 (20 users)	VGG-16 [44]	Biometric Verification	EER = 0.396%
		Own (Middle Quality)	198 (33 users)			EER = 1.275%
		SDUMLA [12] (Low Quality)	636 (106 users)			EER = 3.906%
Huang et al. [45]	Finger vein	Own (Training)	300.000	VGG-16 [44]	Biometric Verification	-
		FVRC2016 - DS1 [46] (Testing)	1000			EER = 0.42%
		FVRC2016 - DS2 [46] (Testing)	1000			EER = 1.41%
		FVRC2016 - DS3 [46] (Testing)	1000			EER = 2.14%
Xie et al. [47]	Finger vein	HKPU [10]	302 (156 users)	LCNN [48]	Biometric Verification	EER = 0.11%
				VGG-16 [44]		EER = 0.12%
Fang et al. [49]	Finger vein	MMCBNU_6000 [22]	600 (100 users)	[50]	Biometric Verification	EER = 0.10%
		SDUMLA [12]	636 (106 users)			EER = 0.47%
Jalilian et al. [51]	Finger vein	UTFVP [42]	360 (60 users)	U-net [52]	Biometric Verification	EER = 4.53%
						EER = 0.41%
				RefineNet [53]		EER = 2.95%
						EER = 1.80%
		SDUMLA [12]	636 (106 users)	SegNet [54]		EER = 2.54%
						EER = 2.17%
Kim et al. [55]	Finger vein & Finger-shape	SDUMLA [12]	636 (106 users)	ResNet-50 [56]	Biometric Verification	EER = 2.34%
		HKPU [10]	302 (156 users)	ResNet-101 [56]		EER = 0.79%
Wang et al. [57]	Hand-dorsal Vein	Own	200 (200 users)	VGG-16 [44]	Biometric Verification	EER = 0.06%
Zhang et al. [58]	Palm Vein	Own	600 (300 users)	Inception ResNet v1 [59]	Biometric Verification	EER = 2.74%

from the merging of two templates. An approach for finger-vein-based biometric verification using CNN and supervised discrete hashing (SDH) has been proposed in [47], where different CNN architectures, such as a light CNN (LCNN) and a modified version of VGG-16, have been fed with pairs of vein images. The SDH scheme is also investigated to improve the performance and to reduce the template size. Fang *et al.* [49] have exploited a lightweight deep-learning framework for finger vein verification. Mini-ROIs from the original image are extracted, based on the evaluation of the adopted network, and both the original image and the mini-ROIs are integrated through a two-stream network. Jalilian *et al.* [51] have used three different fully CNN (FCN) architectures, inspired by the U-net [52], RefineNet [53], and the SegNet [54] networks, to extract the finger vein patterns from NIR finger images. The problem of efficient training and configuration settings for the employed networks has also been taken into account by training the considered FCN architectures with a varying number of manually- and automatically-generated labeled images. Wang *et al.* [57] have proposed a hand-dorsal vein recognition system employing VGG-16, pre-trained on a large-scale database, as a universal feature extractor. A task-specific selective convolutional features (SCF) model, based on spatial weighting, has been proposed to obtain the discriminative features. In addition, spatial pyramid pooling (SPP) has been introduced to obtain the final feature representation. Kim *et al.* [55] have proposed a multimodal biometric recognition system based on finger veins and finger shapes, exploiting CNNs to extract features from the acquired images and compute classification scores. More in detail, the authors have compared the recognition performance of different CNN configurations, namely ResNet-50 and ResNet-101 [56], when fed with the finger vein image or the spectrogram of the finger ROI. They have shown that the performance can be improved by applying score-level fusion approaches. The ResNet architecture has also been exploited by Zhang *et al.* [58] in the framework of a palm-vein-based verification system. The authors have applied a modified version of the Inception ResNet-v1 DNN to extract features, later used for recognition purposes.

IV. DESIGNED FINGER VEIN IDENTIFICATION PIPELINE

The architecture of the proposed identification system is sketched in Figure 1. Its building blocks, that is, data acquisition hardware, preprocessing, and classification modules, are detailed in the following subsections.

A. Data Acquisition Hardware

Our goal is to design a finger vein acquisition system allowing users to donate their finger vein patterns while walking, passing their hands over the acquisition sensor and without any contact, as depicted in principle in Figure 2. Specifically, our system is composed of four PiNoIR-V2 CMOS cameras, equipped with Sony IMX219 8-megapixel sensors [61] having a NIR sensitivity in the wavelength range $400nm - 1000nm$, each driven by a Raspberry PI-V2 card. CMOS-based cameras have been widely used in the literature for vein-pattern acquisition, essentially because of their lower cost with respect to CCD-based cameras which, on the other

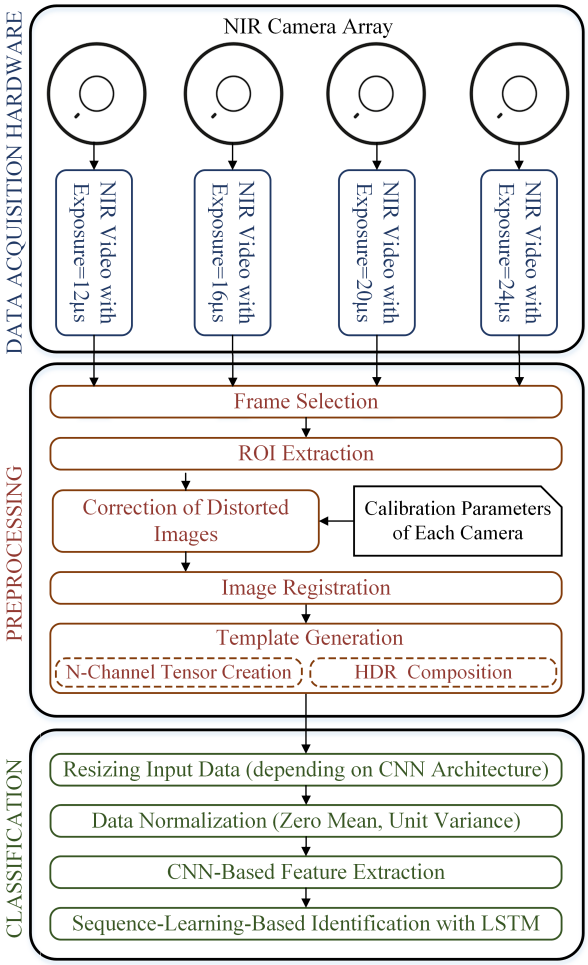

Fig. 1. Representation of the proposed acquisition and processing pipeline.

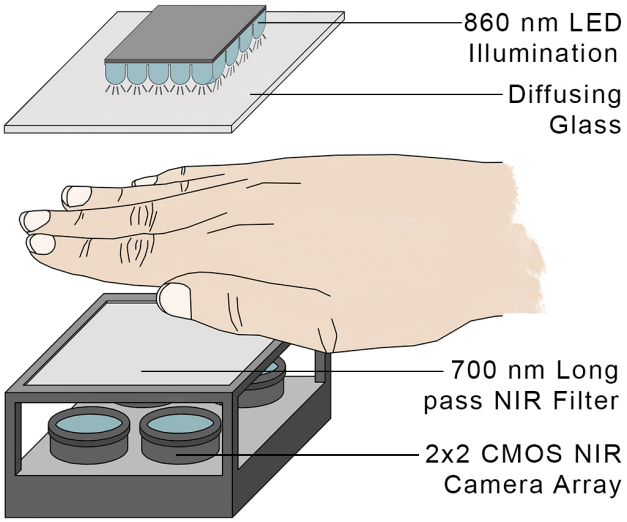

Fig. 2. Graphic representation of the proposed acquisition system.

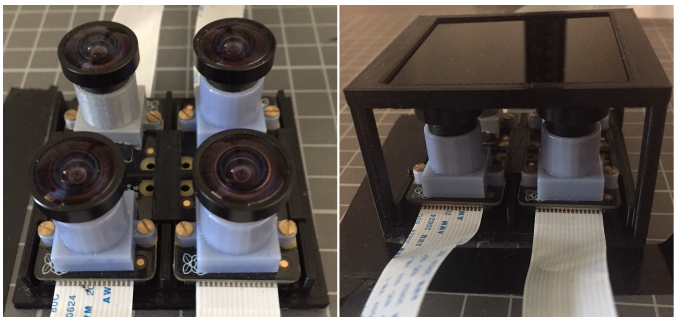

Fig. 3. 2x2 camera array (left), NIR filter on the camera array (right).

side, guarantee higher performance in terms of signal-to-noise ratio (SNR) of the acquired images, especially in the NIR field.

The cameras are arranged in a 2x2 matrix configuration that minimizes the parallax effect in the image acquisition process, with respect to the other camera configurations we have tested. Each of the four cameras employs a different exposure time, namely 12μs, 16μs, 20μs, and 24μs. As outlined in the next section, the use of multiple cameras at different exposures allows mitigating the effects of the adverse conditions characterizing the considered acquisition scenario, which may affect the quality of the captured images. In addition, a $700nm$ longpass NIR filter is placed over the camera array to cut out visible light. The acquisition system is shown in Figure 3.

The employed illuminator is composed of 20 LEDs, arranged in a 5×4 rectangular grid. Specifically, we have employed Osram Opto SFH 4356-UV model IR LEDs with: *i)* dome shaped lens, *ii)* $80mW$ radiant flux, *iii)* $860nm$ peak wavelength, and *iv)* $850nm$ centered wavelength. It is worth mentioning that most of the systems available in the literature use LEDs operating in the range $[850, 930]$nm. As a matter of fact, we have tested LEDs at $830nm$, $850nm$, and $910nm$. The $850nm$ LEDs have proven to be the best-performing in our system. The illuminator is fed with $400mA$ current and $12V$ voltage. A $3mm$-thick white diffusion glass is placed between the hand and the lighting LEDs to obtain a uniform light diffusion over the fingers. The acquisition protocol requires the user to pass the hand between the cameras and the illuminator, thus using a transmission modality.

As mentioned in Section I, it is worth remarking that, in conventional vein-based biometric systems, users have to keep their fingers or hands still during acquisition, and often place them on a support. On the contrary, in our novel approach, we release this constraint and allow the users passing their hands over the sensor as shown in Figure 4, thus significantly improving user convenience and system throughput. During the acquisition, we collect four videos, at the rate of 12 frames per second (fps), each with different exposure times.

B. Preprocessing

After acquisition, the preprocessing steps, sketched in Figure 1, and described hereafter, are performed.

- **Frame selection**: Passing a hand over the acquisition system may require from one to three seconds, depending on the user's behavior. During this time, each of the four employed cameras records up to 36 frames, out of which 9 frames, containing all the hand's fingers, are selected for further processing. Specifically, the frame with the overall lower average luminance across the captured videos is chosen as reference and assumed to be the hand in central position with respect to the device. The four frames before and the four frames after it are then selected for each of the four acquired videos;

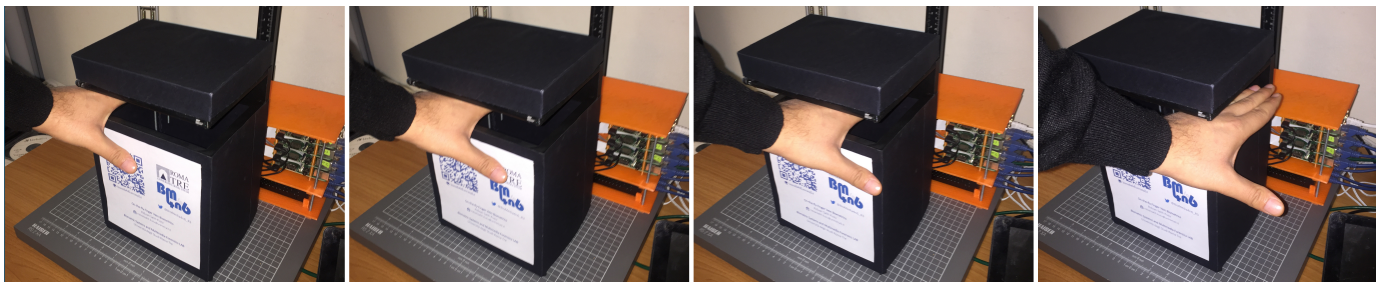

Fig. 4. Acquisition process.

Fig. 5. LDR finger vein templates of a subject acquired using 12μs, 16μs, 20μs, and 24μs exposure times (first four images on the left), and resulting tone-mapped HDR vein template (image on the right). Images contrast enhanced for visualization purposes.

- **ROI extraction**: A ROI of 720x640 (WxH) pixels is extracted from each image. This choice, guarantees that the whole image of the hand is selected, as shown in Figure 5. The so-obtained images have been corrected by using a camera-calibration approach to compensate for fish-eye distortion;
- **Image registration**: Images acquired from different cameras are misaligned due to the parallax effect, caused by the non-negligible size of the employed cameras. Therefore, image registration is needed. In our approach, we have applied the multimodal intensity-based image registration technique proposed in [62];
- **Template generation**: The images extracted from the acquired videos can be affected by low contrast, due to the difficulty in controlling the employed NIR illumination when capturing moving hands, as well as by blur effect due to hand movement. To mitigate these problems, we have applied two different approaches:
  - Use of HDR imaging techniques [63]: the images captured at different exposure times are fused into a single HDR image which does not suffer from under- or over-exposure issues. The generated HDR content can be then converted into a low dynamic range (LDR) image through the use of a tone mapper. In this study we have applied the iCAM06 tone mapping operator [64], due to its superior performance within the proposed framework [65];
  - Use of a 4-channel tensor: the four images, acquired by the four cameras, at the same time but at different exposures, are represented through a single structure. Specifically, four 1-channel grayscale images, representing the luminance content at different exposures, aregrouped to build a 4-channel image tensor.

C. Classification

As mentioned in Section IV-B, during each acquisition, 9 frames are taken from each of the 4 employed cameras, for a total of 36 frames. A single pass of a hand over the sensor therefore generates data characterized by specific spatial behavior, given by the properties of the vein patterns of the four fingers captured in each image, as well as by a temporal behavior, represented by the sequences of frames taken at consecutive instants.

To take advantage of the collected information, we have resorted to DNNs. Specifically, a CNN has been designed to extract reliable features from each processed image. In addition, an RNN has been used to exploit the availability of multiple frames in the acquired videos. In more detail, a long short-term memory (LSTM) network has been exploited to model the observed temporal course of the hand movement.

The proposed CNN-LSTM architecture is discussed in Section V, and its effectiveness is tested in Section VI against state-of-the-art finger vein biometric recognition systems employing CNNs. It is worth pointing out that our approach is multimodal in many aspects:

i) Multiple sensors for the same biometric trait: the same finger vein pattern is acquired using four cameras with different acquisition parameters;
ii) Multiple units of the same biometric trait: four fingers, as a whole, are used together for template generation;
iii) Multiple biometric traits: both finger veins and finger shape are intrinsically exploited in our system.

V. CNN-LSTM Architecture

The proposed finger-vein-based identification system is depicted in Figure 6. Specifically, a sequence of 9 finger vein images, extracted from the recorded video as explained in Section IV-B, is fed into the classifier after resizing and normalizing to zero mean and unit variance. Each element of

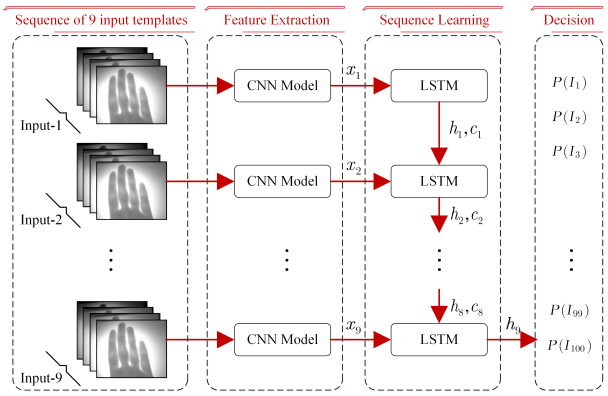

Fig. 6. CNN-LSTM architecture of the proposed system.

the sequence is processed separately by the same CNN model to create visual features for each element of the sequence. Then, an LSTM network, able to find and keep trace of the temporal dependencies within the input feature sequence [66], is applied to predict the most probable identity to be associated with the query acquisition. The proposed neural network architecture is discussed in detail in the two following sections, which describe the used CNN and LSTM topologies.

A. Proposed CNN Topology

The architecture of the CNN adopted in the proposed system, shown in Figure 7, has been specifically designed for finger-vein-based recognition tasks. The configuration here described in terms of layers, kernel sizes, and number of features produced by the fully-connected layers, has been obtained by maximizing the recognition performance with respect to the network parameters. The approximately-rectangular shape of the fingers has been taken into account when selecting the configurations to be tested, thus using rectangular kernels instead of squared ones, commonly used in standard CNNs.

The network proposed in this paper, namely Vein-CNN (V-CNN), has 6 convolutional layers, 4 max-pooling layers, 2 fully-connected layers, and an output layer, with 13 layers in total, as summarized in Table III. Specifically:

- *IN*: the input layer receives $[320 \times 360 \times N_c]$ data, with finger vein images of size $[320 \times 360]$ and N_c channels for each image. As described in Section IV-B, we may have either $N_c = 1$ or $N_c = 4$ in the proposed configurations, while a resizing is needed to obtain finger vein images of the required input size, as described in Section IV-B;
- *Group-1 ($CL_1 - M_1$)*: the first hidden layer group is composed of 64 convolutional filters with size $[3 \times 3 \times 64]$, followed by a batch normalization, a rectifier linear unit (ReLU), and a max-pooling layer of size $[2 \times 2]$. After the convolution and down-sampling, $[160 \times 180 \times 64]$ low-level features are extracted from the input data;
- *Group-2 ($CL_2 - CL_3 - M_2$)*: the second hidden layer group consists of two layers with 128 convolutional filters, each followed by a batch normalization and a ReLU. Kernel size for the former one is set at $[11 \times 5]$ to capture patterns along the finger orientation, mainly vertical. Features of $[80 \times 90 \times 128]$ dimensions are obtained after 2-layered convolution and down-sampling with $[2 \times 2]$ max-pooling;
- *Group-3 ($CL_4 - CL_5 - M_3$)*: the third hidden layer group consists of two layers with 256 convolutional filters, each followed by a batch normalization and a ReLU. Differently from the previous group, the kernel size for the first convolutional layer is set at $[5 \times 11]$ to capture patterns across fingers. Features of $[40 \times 45 \times 256]$ dimensions are obtained after 2-layered convolution and down-sampling with $[2 \times 2]$ max-pooling;
- *Group-4 ($CL_6 - M_4$)*: the fourth hidden layer group is composed of 512 convolutional filters with size $[3 \times 3 \times 512]$, followed by a batch normalization, a ReLU, and a max-pooling layer of size $[5 \times 5]$. After convolution and down-sampling, the output of previous group is transformed into $[8 \times 9 \times 512]$ features;
- *Group-5 ($FC_1 - FC_2$)*: the fifth hidden layer group contains two fully-connected layers, each followed by a ReLU, and a dropout regularization. The 50% of hidden units are dropped to reduce overfitting while training. This group generates a $[1024 \times 1 \times 1]$ feature map x;
- *OUT*: the output layer consists of N_I neurons, where N_I is the number of unique identities/subjects in the database. A fully-connected softmax layer is employed at this stage, giving as output the "1-to-N_I" match probabilities for the considered subjects.

As already mentioned, the described network has been specifically designed to deal with the finger vein images generated in the proposed system. In order to test its effectiveness, in Section VI its behavior is compared against three different state-of-the-art CNN architectures, namely VGG-19 [44], Densenet-201 [67], and Inception-v3 [68]. These networks have been selected due to their high performance when tested within the ImageNet Large Scale Visual Recognition Challenge framework [69], [70].

B. LSTM Topology

Long short-term memory (LSTM) networks represent the most-commonly-employed kind of RNN that process data in sequential order and keep their hidden state through time. Although different implementations of LSTM units exist, they generally consist of a memory cell (c_t), an input gate (i_t), an output gate (o_t), and a forget gate (f_t). While the cell is responsible for storing information values over arbitrary time intervals, the gates regulate the cell input and output information flows over time. The capabilities of the gates are synthesized in Figure 8:

- The input gate manages the process of adding new information into the cell from the sequence flow;
- The forget gate deletes from the cell the information that is no longer necessary for the LSTM unit;
- The output gate selects the relevant information to be given as output.

More in detail, letting $\sigma(x) = (1 + \mathrm{e}^{-x})^{-1}$ be the *sigmoid* function and $\phi(x) = 2\sigma(2x) - 1$ be the *hyperbolic tangent* function, the LSTM unit is updated at time-step t as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

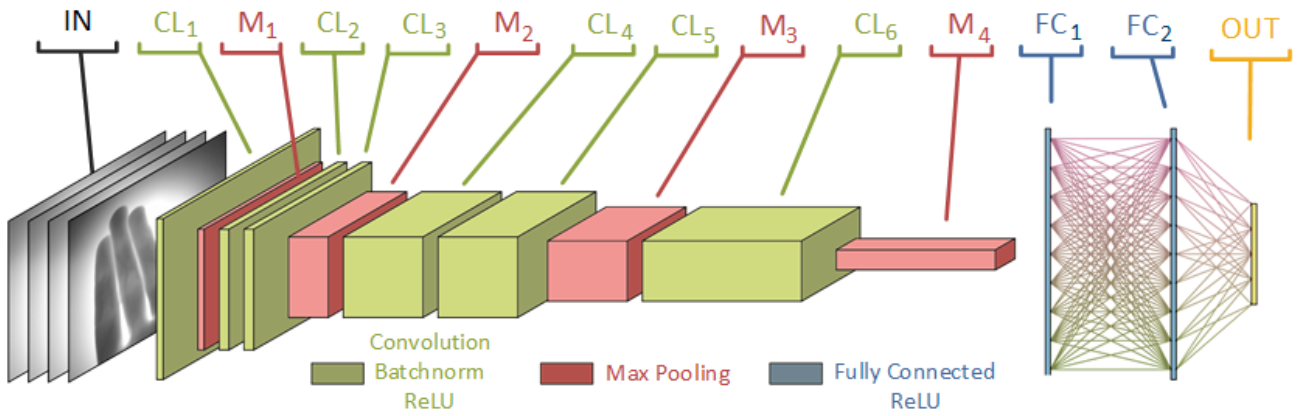

Fig. 7. V-CNN Architecture.

TABLE III
PROPOSED CNN CONFIGURATION (V-CNN).

Abbreviation	Layer Type	Number of Filter	Size of Feature Map	Size of Kernel	Number of Stride	Number of Padding
IN	Image Input Layer	-	$320 \times 360 \times N_c$	-	-	-
CL_1	Convolutional Layer-1	64	$320 \times 360 \times 64$	3×3	1×1	1×1
M_1	Max-Pooling Layer-1	1	$160 \times 180 \times 64$	2×2	2×2	0×0
CL_2	Convolutional Layer-2	128	$160 \times 180 \times 128$	11×5	1×1	5×2
CL_3	Convolutional Layer-3	128	$160 \times 180 \times 128$	5×5	1×1	2×2
M_2	Max-Pooling Layer-2	1	$80 \times 90 \times 128$	2×2	2×2	0×0
CL_4	Convolutional Layer-4	256	$80 \times 90 \times 256$	5×11	1×1	2×5
CL_5	Convolutional Layer-5	256	$80 \times 90 \times 256$	5×5	1×1	2×2
M_3	Max-Pooling Layer-3	1	$40 \times 45 \times 256$	2×2	2×2	0×0
CL_6	Convolutional Layer-6	512	$40 \times 45 \times 512$	3×3	1×1	1×1
M_4	Max-Pooling Layer-4	1	$8 \times 9 \times 512$	5×5	5×5	0×0
FC_1	Fully-Connected Layer-1	1024×1				
FC_2	Fully-Connected Layer-2	1024×1				
OUT	Output Layer	$N_I \times 1$				

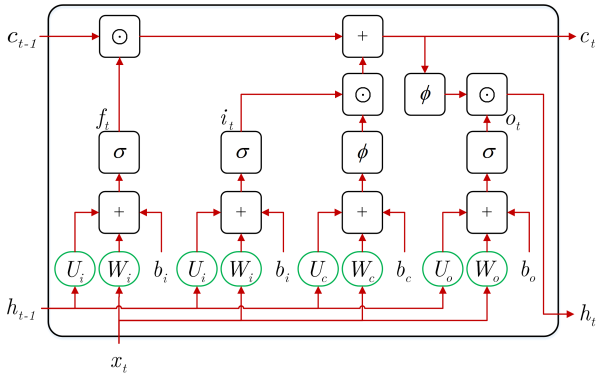

Fig. 8. Diagram for LSTM unit at time-step t.

$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \phi(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \odot \phi(c_t),
\end{aligned}
$$

where $i_t, f_t, o_t \in \mathbb{R}^h$ are the activation vectors for input gate, forget gate, and output gate respectively, $x_t \in \mathbb{R}^d$, with $d = 1024$, is the input vector to the LSTM unit, $h_t \in \mathbb{R}^h$ is the output vector of the LSTM unit, $c_t \in \mathbb{R}^h$ is the cell-state vector, $W \in \mathbb{R}^{h \times d}$, $U \in \mathbf{R}^{h \times h}$, $b \in \mathbb{R}^h$ are the weight matrices and bias vector for the gates to be learned during training [71], and $\odot$ is the Hadamard element-wise product.

In the used LSTM architecture, as illustrated in Figure 6, CNN models are initially trained using every frame of the input sequences as if they were separate still images. After that, each element of an image sequence is converted into a d-dimensional feature vector x using the trained CNN models, and employed as an input vector for an LSTM unit. For instance, after discarding the final CNN layer, i.e., the output layer *OUT*, the proposed V-CNN is able to create a 1024-dimensional feature vector from an input template. A sequence of 9 feature vectors is thus associated to each hand pass, whose temporal behavior is then modeled through a single LSTM layer with $h = 2048$ hidden units. The dimensionality of the employed LSTM feature space has been set according to a trial-and-error approach, choosing the one giving the best recognition accuracy. Letting x_t be the feature vector extracted from a frame of an input sequence, and $P(I)$ the likelihood of unique identities on a query, the prediction of the LSTM model can be represented as $\langle x_1, x_2, \ldots, x_T \rangle \mapsto P(I)$, with $T = 9$ and N_I possible unique identities.

The rational behind the separate training of the proposed CNN and LSTM networks is the need to train the CNN over the largest possible amount of data. The joint training of the two networks would have instead required to significantly lower the number of samples used to estimate the optimal CNN parameters with a consequent worsening of

the recognition performance. Furthermore, this choice allows performing a fair comparison among the V-CNN and the other state-of-the-art networks, considered in the experimental tests described in Section VI, in terms of their effectiveness to extract discriminative features.

C. Network Initialization and Optimization

For all the CNN architectures considered in this study, a cross-entropy (CE) loss function is preferred for back-propagation. A stochastic gradient descent (SGD) with a batch size of 32 is used for training optimization. During the SGD optimization, *i)* learning rate stars at 0.001 and it is divided by 10 after each 30-epoch iteration, *ii)* momentum is set to 0.9 for accelerating the gradients in the right directions, and *iii)* the L_2 regularization penalty (weight decay) is set to 0.01. Since a poor initialization of the neural network weights can divert the learning steps to a wrong path, the following strategies are adopted in our experiments: *i)* random initialization with a normal distribution $\mathcal{N}(0,\, 2/n)$ for convolutional layers, where $n = K_w \times K_h \times N_{C_{out}}$, being K_w, K_h the kernel sizes, and $N_{C_{out}}$ the output channel size, *ii)* unit-weight initialization for batch normalization, and *iii)* random initialization with a normal distribution $\mathcal{N}(0,\, 0.01)$ for fully-connected layers.

For the LSTM architecture, a categorical-hinge (CH) function is preferred for back-propagation and an AMSGrad variant of ADAM, with a batch size of 32, is used for training optimization. The learning rate is set to 0.0001 and divided by 10 after each 30-epoch iteration for LSTM training. During the initialization of the weights on LSTM network, Glorot uniform initializer for input kernels, and orthogonal initializer for the recurrent kernels, are preferred.

The maximum number of training epochs is set to 90 for both CNN and LSTM networks, with the validation loss monitored to determine when to stop the learning processes.

VI. Experimental Tests

In order to evaluate the effectiveness of the proposed CNN-LSTM architecture within the considered framework, we have performed several tests over a database collected at our Institution. Specifically, we have recorded 10 acquisitions, each made by a sequence of 9 finger vein images, from the left hands of $N_I = 100$ subjects, 33 female, and 67 male.

It is worth pointing out that, since the acquisition modality proposed in this paper is novel, neither other datasets nor other methodologies are available for performance comparison. Nonetheless, we have included in our tests the comparison among the identification performance guaranteed by the proposed V-CNN network, when single-image finger vein acquisitions are considered, and the most-commonly-employed CNNs existing in the literature. In more detail, in Section VI-A, the results proving the effectiveness of using several cameras at different exposures, with respect to the use of a single camera, are reported. Then, in Section VI-B we compare the recognition results achievable with the proposed V-CNN framework against those attainable with state-of-the-art convolutional networks. The advantages of exploiting the properties of an LSTM network in addition to a CNN are then detailed in Section VI-C. Preprocessing has been performed using MATLAB™ (R2017b) and PyTorch 0.4.0 has been chosen to build network architectures with a system configuration of 32Gb RAM, NVIDIA™ Titan V graphics card, i7-3.4GHz processors, and Windows™ 10 operating system.

TABLE IV
Mean identification accuracy with single-exposure inputs (E_i) vs. their joint usages as HDR and 4-channel tensor.

Training Acquisitions	$\mathbf{E}_1$ ($12\mu s$)	$\mathbf{E}_2$ ($16\mu s$)	$\mathbf{E}_3$ ($20\mu s$)	$\mathbf{E}_4$ ($24\mu s$)	HDR Input	Tensor Input
2	80.95%	81.21%	80.99%	80.75%	82.62%	90.18%
3	88.76%	90.71%	90.33%	89.35%	91.08%	94.89%
4	92.19%	93.86%	93.14%	93.39%	94.65%	96.61%
5	93.94%	95.71%	95.74%	95.81%	96.16%	97.62%

A. Usage of Multiple-Exposure Finger Vein Images

In the proposed architecture, we further speculate on the advantages achievable when using multiple images taken at different exposures [65]. Specifically, we have compared in Table IV the performance obtained when a single camera is used against the joint usage of the images taken at the four considered exposures. The reported results are referred to the use of only the V-CNN architecture for identification, selecting, for each identity at each iteration, the finger vein images from:

- Five acquisitions for training;
- One distinct acquisition for validation;
- The remaining four acquisitions for testing,

and performing a 5-fold cross-validation. Out of the five acquisitions selected for training, a different number N_T of samples has been selected at each iteration to evaluate the achievable identification performance at increasing size of the training set. From the rank-1 identification accuracy given in Table IV it is possible to confirm the conclusions drawn in [65], noticing that HDR images allow achieving better performance than the usage of singe-exposure images. More interestingly, using the collected data as 4-channel tensors in the proposed V-CNN guarantees further improvements. The results obtained when considering images taken at different exposures also highlight that, besides the shape of the captured fingers which remains the same independently of the exposure, the proposed system exploits the finger vein patterns for recognition purposes. In fact, the use of finger vein images acquired at different exposures impacts the identification performance, as shown in Section IV-C. It is worth remarking that all the experiments have been made without applying enhancement techniques on finger vein inputs, which represents an additional benefit of the proposed approach.

B. CNN-Only Comparisons

To speculate about the performance of the proposed V-CNN architecture, in this section we do not consider the temporal/sequential relations among frames. In more detail, the proposed framework is compared against three different state-of-the-art CNN architectures, namely VGG-19, Densenet-201, and Inception-v3, trained on the collected database as outlined in Section V-C for the proposed V-CNN, with the same cross-validation strategies described in Section VI-A.

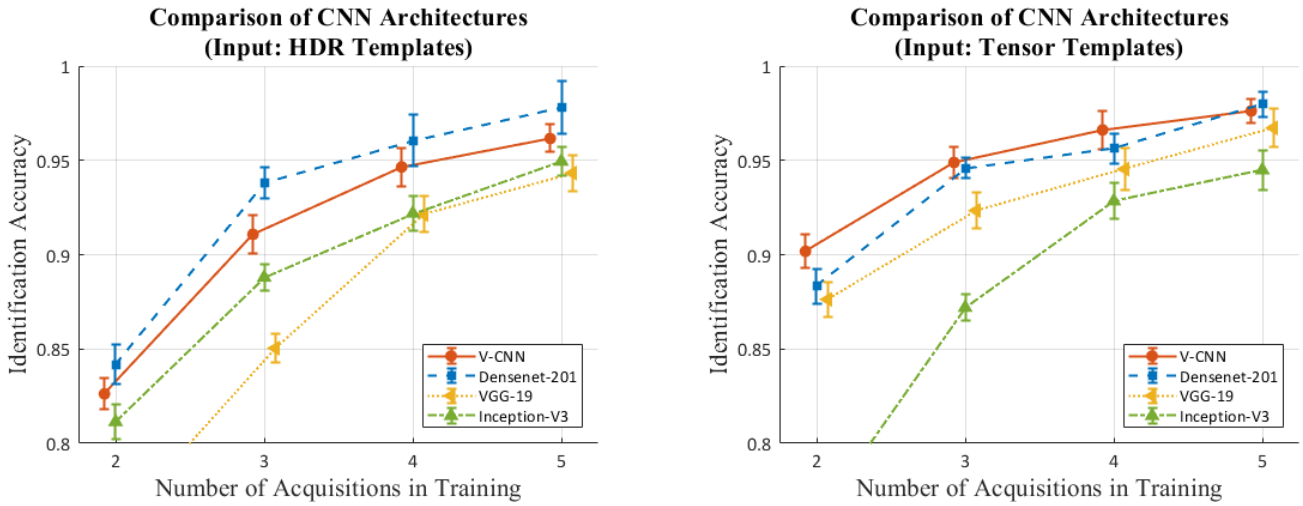

Fig. 9. CNNs performance comparison for different training acquisitions using HDR (left) and 4-channel tensor (right) templates.

TABLE V
CNNs IDENTIFICATION ACCURACY FOR DIFFERENT NUMBERS OF TRAINING ACQUISITIONS USING HDR AND 4-CHANNEL TENSOR TEMPLATES.

Number of Acquisitions	V-CNN		VGG-19		Inception-V3		Densenet-201	
	HDR Input	Tensor Input	HDR Input	Tensor Input	HDR Input	Tensor Input	HDR Input	Tensor Input
2	$82.62 \pm 0.84\%$	$90.18 \pm 0.87\%$	$76.31 \pm 2.04\%$	$87.63 \pm 0.90\%$	$81.14 \pm 0.91\%$	$76.01 \pm 1.45\%$	$84.18 \pm 1.04\%$	$88.32 \pm 0.90\%$
3	$91.08 \pm 1.02\%$	$94.89 \pm 0.82\%$	$85.04 \pm 0.76\%$	$92.34 \pm 0.93\%$	$88.80 \pm 0.69\%$	$87.20 \pm 0.72\%$	$93.79 \pm 0.82\%$	$94.60 \pm 0.56\%$
4	$94.65 \pm 1.03\%$	$96.61 \pm 1.02\%$	$92.13 \pm 0.95\%$	$94.54 \pm 1.09\%$	$92.19 \pm 0.91\%$	$92.84 \pm 0.96\%$	$96.05 \pm 1.36\%$	$95.64 \pm 0.80\%$
5	$96.16 \pm 0.79\%$	$97.62 \pm 0.66\%$	$94.33 \pm 0.94\%$	$96.72 \pm 1.00\%$	$94.95 \pm 0.78\%$	$94.49 \pm 1.03\%$	$97.80 \pm 1.39\%$	$97.97 \pm 0.68\%$

Figure 9 shows that, when 4-channel tensor inputs are exploited, the accuracy of the proposed V-CNN architecture is higher than the one of the other networks. On the other hand, if HDR content is considered as input, Densenet-201 outperforms the proposed V-CNN, which is however better than both VGG-19 and Inception-v3. As shown in Table V, the use of the 4-channel tensors as input significantly improves the identification accuracy, especially when a limited number of acquisitions is available for training. For instance, when 2 samples are used for training, the observed improvements with respect to the use of HDR inputs are 7.56%, 11.32%, and 4.14% for V-CNN, VGG-19, and Densenet-201 respectively. Accordingly, the highest positive impact of using 4-channel tensor templates is observed when using the VGG-19 network.

In summary, when comparing V-CNN with other state-of-the-art CNNs, the highest identification accuracy scores are obtained when 4-channel input tensors are considered.

C. Exploitation of the Temporal Information

The use of the on-the-fly acquisition protocol allows recording also temporal information regarding how the hand is passed over the sensor. Actually, the LSTM network described in Section V-B has been designed to exploit the temporal evolution of the discriminative features during each acquisition.

In order to show the effectiveness of our approach, we have considered two alternative methods for fusing the spatial information derived from multiple frames, yet without exploiting any temporal information. Specifically, we have implemented a score-level fusion (SF) as well as a decision-level fusion (DF) strategy over the features extracted by the CNN processing individual frames. In more detail, SF is performed by averaging the likelihoods obtained as predictions from the CNN models for each of the nine separate frames of an acquisition. Majority voting is instead performed to implement DF once the predictions for each frame are provided by the CNN. The improvements achieved using SF and DF over the acquired sequences of images are shown in Figure 10. If we consider the results reported in Table V as a reference, both SF and DF increase the accuracy of 2%-5%, when considering two acquisitions for training. Nevertheless, the addition of more training samples slightly reduces the relevance of the improvement on the identification accuracy. Moreover, both SF and DF result in similar patterns over the identification performance, which means that the two approaches are not significantly different from each other ($p-$value $= 0.416$ in terms of paired t-tests).

The effects of exploiting the temporal information through the proposed CNN-LSTM framework are shown in Figure 11. The obtained results are also reported in Table VI, where the accuracy achieved by the best-performing CNNs, that is, V-CNN and Densenet-201, are reported. Moreover, all the considered CNN architectures are also compared in terms of 1-tailed t-test, to evaluate if the differences in the obtained results are statistically significant when a limited number of acquisitions are used for training. As shown in Table VII, the proposed CNN-LSTM network is able to exploit the temporal behavior of the hand movement better than simple strategies based on SF or DF. In more detail, the effects of LSTM are remarkably higher than those of SF and DF, where the p-values are less than 0.05 for all comparisons.

Restricting our analysis to the best-performing CNNs, that is, V-CNN and Densenet-201, Densenet-201+LSTM has better performance than V-CNN+LSTM with HDR input templates. However, the performance of Densenet-201+LSTM is not

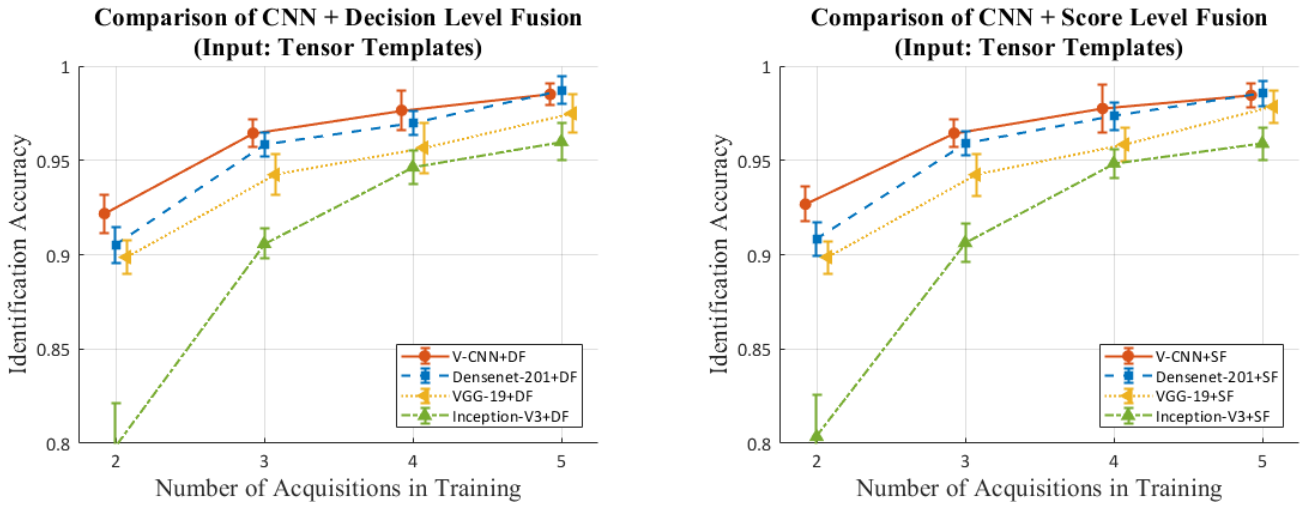

Fig. 10. Performance comparison of simple fusion techniques based on DF (left) and SF (right) over CNN-based feature extracted from frames of a sequence, for different training acquisitions and using 4-channel input templates.

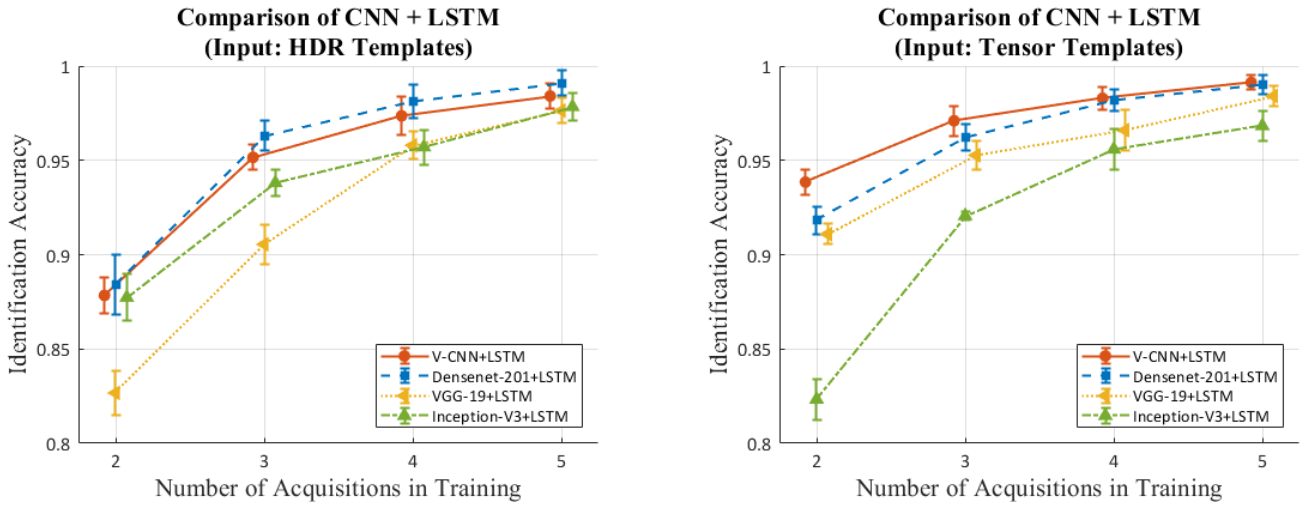

Fig. 11. Performance comparison of fusion techniques based on LSTM over CNN-based feature extracted from frames of a sequence, for different training acquisitions and using HDR (left) and 4-channel (right) input templates.

TABLE VI
MEAN IDENTIFICATION ACCURACY FOR FUSION TECHNIQUES BASED ON DF, SF, AND LSTM, OVER V-CNN AND DENSENET-201 FEATURES.

Number of Training Acquisitions	V-CNN		V-CNN+DF		V-CNN+SF		V-CNN+LSTM	
	HDR Input	Tensor Input	HDR Input	Tensor Input	HDR Input	Tensor Input	HDR Input	Tensor Input
2	82.62%	90.18%	86.22%	92.18%	86.22%	92.68%	87.84%	93.85%
3	91.08%	94.89%	93.58%	96.44%	93.78%	96.44%	95.16%	97.10%
4	94.65%	96.61%	96.54%	97.65%	96.94%	97.75%	97.36%	98.29%
5	96.16%	97.62%	97.94%	98.50%	98.04%	98.45%	98.41%	99.13%
Number of Training Acquisitions	**Densenet-201**		**Densenet-201+DF**		**Densenet-201+SF**		**Densenet-201+LSTM**	
	HDR Input	Tensor Input	HDR Input	Tensor Input	HDR Input	Tensor Input	HDR Input	Tensor Input
2	84.18%	88.32%	86.97%	90.53%	87.17%	90.83%	88.43%	91.81%
3	93.79%	94.60%	95.44%	95.84%	95.59%	95.89%	96.31%	96.23%
4	96.05%	95.64%	97.15%	96.99%	97.30%	97.34%	98.12%	98.19%
5	97.80%	97.97%	98.45%	98.70%	98.55%	98.55%	99.10%	99.02%

significantly better when the number of acquisitions used in the training stage is 2 and 4, as summarized in Table VIII. Instead, V-CNN+LSTM outperforms the other network architectures when using 4-channel tensor inputs, guaranteeing the best rank-1 identification results. It is worth mentioning that, due to its higher complexity, Densenet-201 requires approximately 20% more time than V-CNN to be trained.

Tensor inputs are therefore taken into account to show the cumulative match characteristic (CMC) curves achievable with the proposed CNN-LSTM finger-vein-based identification system when considering limited enrolment data, namely two or three acquisitions, for each subject. Figure 12 illustrates the obtained results, showing that V-CNN-LSTM allows achieving better performance than the other architectures.

TABLE VII
SIGNIFICANCE (p-VALUES) OF LSTM VS. SF AND DF PERFORMANCE COMPARISON, WITH RESPECT TO 1-TAILED T-TEST WHEN $N_T = 2$.

		LSTM			
		V-CNN	**Densenet-201**	**VGG-19**	**Inception-V3**
Tensor Input	**SF**	0.003	0.014	0.002	0.031
	DF	0.007	0.002	0.008	0.026
HDR Input	**SF**	0.008	0.049	0.032	0.012
	DF	0.010	0.031	0.046	0.001

TABLE VIII
SIGNIFICANCE (p-VALUES) OF V-CNN+LSTM VS. DENSENET-201+LSTM PERFORMANCE COMPARISON, FOR 1-TAILED T-TESTS (SIGNIFICANT p-VALUES LOWER THAN 0.05 SHOWN IN BOLD).

	Number of Training Acquisitions			
	2	**3**	**4**	**5**
Tensor Input	**0.012**	**0.044**	0.389	0.179
HDR Input	0.224	**0.014**	0.141	**0.019**

To sum up, among all the CNN architectures evaluated in this study, V-CNN and Densenet-201 perform better than the others. As summarized in Table VI, the use of these networks followed by an LSTM model produces the best improvements when compared to the baseline results. As an example, the use of DF on 4-channel tensors with two training acquisitions gives 92.18% and 90.53% accuracy rates for V-CNN and Densenet-201, respectively, while the identification results increase to 93.85% and 91.81% when adding the LSTM module. The highest rank-1 accuracy achieved in our experiments is 99.13%, obtained when the proposed V-CNN+LSTM model is trained with five acquisitions, represented as 4-channel tensor template sequences.

VII. CONCLUSIONS

In this paper we have proposed, for the first time in the literature, an innovative on-the-fly finger-vein-based biometric recognition system that allows a user being identified while passing a hand over the sensor without requiring any contact. The acquisition module has been built using low-cost sensors and it has been designed to allow free hand movement, with consequent high user convenience during both enrolment and recognition. Multiple cameras with different exposure times, also capturing the dynamic movement of the hand over the sensors, have been used, and a database comprising on-the-fly hand acquisitions from 100 subjects has been collected. The proposed approach exploits both the images acquired at different exposure times and the temporal behavior of the moving hand over the sensors. Deep learning approaches have been used in both scenarios.

Our analysis shows that the use of multiple-exposure data increases the recognition accuracy with respect to the use of single-exposure images, and that the exploitation of multi-channel LDR images taken at different exposure times, as raw input templates, leads to further improvements in the identification accuracy. We have also proposed a novel CNN architecture, namely V-CNN, customized for finger vein identification. Despite its simplicity, the proposed V-CNN outperforms other state-of-the-art CNN architectures. In addition, for the first time in the literature, we have exploited the temporal information related to the hand movement over the sensor, and we have shown that when CNN topologies are used for feature extraction, and LSTM networks are fed by the sequential features based on hand behaviour, a significant identification accuracy improvement is observed.

This work paves the road to further research in the field of on-the-fly finger vein recognition methods. The analysis of open-set and verification scenarios, presentation attack detection, computational complexity reduction, as well as hardware and software optimization, are just few examples of the challenges worth to be tackled in the near future.

VIII. ACKNOWLEDGEMENTS

This work has been partially supported by the EU Horizon 2020 Framework for Research and Innovation under Grant Agreement Number 675087 as part of the AMBER (enhAnced Mobile BiomEtRics) Marie Sklodowska-Curie project and by the Italian PRIN project COntactlesS Multibiometric mObile System in the wild: COSMOS.

REFERENCES

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[2] S. Marcel, M. Nixon, F. J., and N. Evans, *Handbook of Biometric Anti-Spoofing - Presentation Attack Detection*. Springer, 2019.

[3] R. Das, E. Piciucco, E. Maiorana, and P. Campisi, "Convolutional Neural Network for Finger-Vein-Based Biometric Identification," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 360–373, 2019.

[4] Y. Zhou and A. Kumar, "Human Identification Using Palm-Vein Images," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 4, pp. 1259–1274, 2011.

[5] W. Kang and Q. Wu, "Contactless Palm Vein Recognition Using a Mutual Foreground-Based Local Binary Pattern," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1974–1985, 2014.

[6] J. E. S. Pascual, J. Uriarte-Antonio, R. Sanchez-Reillo, and M. G. Lorenz, "Capturing Hand or Wrist Vein Images for Biometric Authentication Using Low-Cost Devices," in *6th IEEE International. Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, 2010.

[7] C.-L. Lin and K.-C. Fan, "Biometric Verification Using Thermal Images of Palm-Dorsa Vein Patterns," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 199–213, 2004.

[8] J. Wang and G. Wang, "Hand-Dorsa Vein Recognition with Structure Growing Guided CNN," *Optik-International Journal for Light and Electron Optics*, vol. 149, pp. 469–477, 2017.

[9] A. Uhl, C. Busch, S. Marcel, and R. Veldhuis, *Handbook of Vascular Biometric*. Springer, 2020.

[10] A. Kumar and Y. Zhou, "Human Identification Using Finger Images," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2228–2244, 2012.

[11] H. T. Van, T. T. Thai, and T. H. Le, "Robust Finger Vein Identification Base on Discriminant Orientation Feature," in *7th International Conference on Knowledge and Systems Engineering (KSE)*, 2015.

[12] Y. Yin, L. Liu, and X. Sun, *SDUMLA-HMT: A Multimodal Biometric Database*. Springer Berlin Heidelberg, 2011.

[13] W. Jia, D.-S. Huang, and D. Zhang, "Palmprint Verification Based on Robust Line Orientation Code," *Pattern Recognition*, vol. 41, no. 5, pp. 1504 – 1513, 2008.

[14] Y. Lu, S. J. Xie, S. Yoon, and D. S. Park, "Finger Vein Identification Using Polydirectional Local Line Binary Pattern," in *International Conference on ICT Convergence (ICTC)*, pp. 61–65, Oct 2013.

[15] T. S. Ong, J. H. Teng, K. S. Muthu, and A. B. J. Teoh, "Multi-Instance Finger Vein Recognition Using Minutiae Matching," in *6th International Congress on Image and Signal Processing (CISP)*, vol. 03, pp. 1730–1735, Dec 2013.

[16] S. Qiu, Y. Liu, Y. Zhou, J. Huang, and Y. Nie, "Finger-Vein Recognition Based on Dual-Sliding Window Localization and Pseudo-Elliptical Transformer," *Expert Systems with Applications*, vol. 64, pp. 618 – 632, 2016.

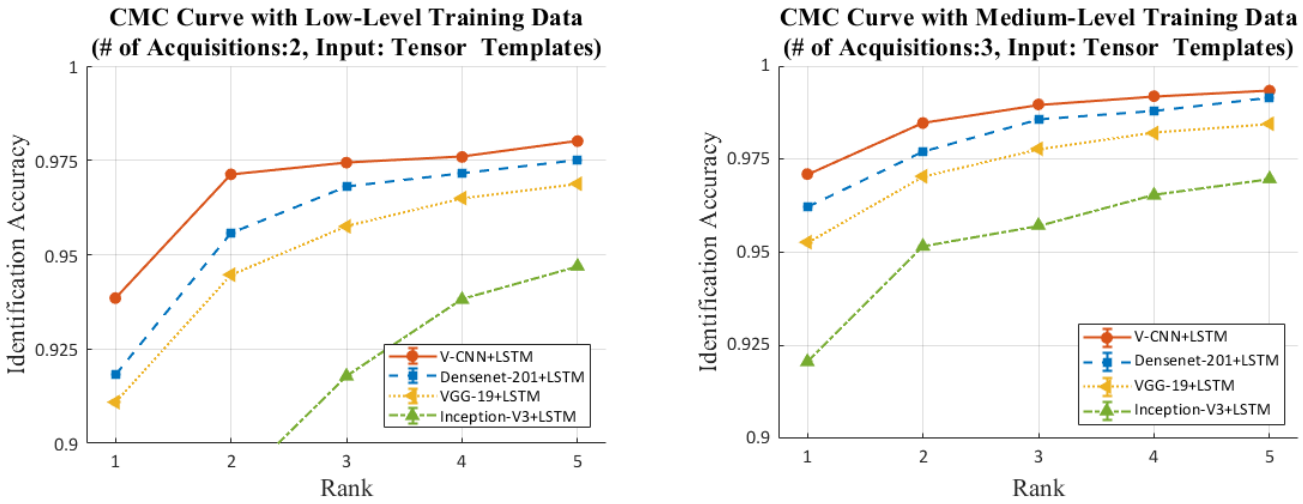

Fig. 12. CMC curves with 4-channel input tensors under the assumption that only two or three acquisitions are collected for the user enrolment.

[17] M. S. M. Asaari, S. A. Suandi, and B. A. Rosdi, "Fusion of Band Limited Phase Only Correlation and Width Centroid Contour Distance for Finger Based Biometrics," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3367–3382, 2014.

[18] S. J. Xie, S. Yoon, J. Yang, Y. Lu, D. S. Park, and B. Zhou, "Feature Component-Based Extreme Learning Machines for Finger Vein Recognition," *Cognitive Computation*, vol. 6, pp. 446–461, Sep 2014.

[19] A. Banerjee, S. Basu, S. Basu, and M. Nasipuri, "ARTeM: A New System for Human Authentication Using Finger Vein Images," *Multimedia Tools and Applications*, Mar 2017.

[20] L. Yang, G. Yang, Y. Yin, and X. Xi, "Finger Vein Recognition with Anatomy Structure Analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1892–1905, 2018.

[21] L. Yang, G. Yang, X. Xi, K. Su, Q. Chen, and Y. Yin, "Finger Vein Code: From Indexing to Matching," *IEEE Transactions on Information Forensics and Security*, 2018.

[22] Y. Lu, S. J. Xie, S. Yoon, Z. Wang, and D. S. Park, "An Available Database for the Research of Finger Vein Recognition," in *International Conference on Image and Signal Processing (CISP)*, vol. 1, pp. 410–415, IEEE, Dec 2013.

[23] N. Miura, A. Nagasaka, and T. Miyatake, "Feature Extraction of Finger-Vein Patterns Based on Repeated Line Tracking and its Application to Personal Identification," *Machine Vision and Applications*, vol. 15, no. 4, pp. 194–203, 2004.

[24] N. Miura, A. Nagasaka, and T. Miyatake, "Extraction of Finger-Vein Patterns Using Maximum Curvature Points in Image Profiles," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 8, pp. 1185–1194, 2007.

[25] W. Song, T. Kim, H. C. Kim, J. H. Choi, H.-J. Kong, and S.-R. Lee, "A Finger-Vein Verification System Using Mean Curvature," *Pattern Recognition Letters*, vol. 32, no. 11, pp. 1541–1547, 2011.

[26] C.-B. Yu, H.-F. Qin, Y.-Z. Cui, and X.-Q. Hu, "Finger-Vein Image Recognition Combining Modified Hausdorff Distance with Minutiae Feature Matching," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 1, pp. 280–289, Dec 2009.

[27] F. Liu, G. Yang, Y. Yin, and S. Wang, "Singular Value Decomposition Based Minutiae Matching Method for Finger Vein Recognition," *Neurocomputing*, vol. 145, pp. 75 – 89, 2014.

[28] L. Wang, G. Leedham, and D. S.-Y. Cho, "Minutiae Feature Analysis for Infrared Hand Vein Pattern Biometrics," *Pattern Recognition*, vol. 41, no. 3, pp. 920–929, 2008.

[29] J.-D. Wu and C.-T. Liu, "Finger-Vein Pattern Identification Using SVM and Neural Network Technique," *Expert Systems with Applications*, vol. 38, no. 11, pp. 14284–14289, 2011.

[30] J.-D. Wu and C.-T. Liu, "Finger-Vein Pattern Identification Using Principal Component Analysis and the Neural Network Technique," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5423–5427, 2011.

[31] Z. Liu, Y. Yin, H. Wang, S. Song, and Q. Li, "Finger Vein Recognition with Manifold Learning," *Journal of Network and Computer Applications*, vol. 33, no. 3, pp. 275–282, 2010.

[32] F.-X. Guan, K. Wang, J. Liu, and H. MA, "Bi-Direction Weighted (2D) 2 PCA with Eigenvalue Normalization One for Finger Vein Recognition," *Pattern Recognition and Artificial Intelligence*, vol. 24, no. 3, pp. 417–424, 2011.

[33] G. Yang, X. Xi, and Y. Yin, "Finger Vein Recognition Based on (2D) 2 PCA and Metric Learning," *BioMed Research International*, vol. 2012, 2012.

[34] E. C. Lee, H. C. Lee, and K. R. Park, "Finger Vein Recognition Using Minutia-Based Alignment and Local Binary Pattern-Based Feature Extraction," *International Journal of Imaging Systems and Technology*, vol. 19, no. 3, pp. 179–186, 2009.

[35] B. A. Rosdi, C. W. Shing, and S. A. Suandi, "Finger Vein Recognition Using Local Line Binary Pattern," *Sensors*, vol. 11, no. 12, pp. 11357–11371, 2011.

[36] E. C. Lee, H. Jung, and D. Kim, "New Finger Biometric Method Using Near Infrared Imaging," *Sensors*, vol. 11, no. 3, pp. 2319–2333, 2011.

[37] X. Li, S. Guo, F. Gao, and Y. Li, "Vein Pattern Recognitions by Moment Invariants," in *1st International Conference on Bioinformatics and Biomedical Engineering*, pp. 612–615, IEEE, 2007.

[38] J. Peng, N. Wang, A. A. A. El-Latif, Q. Li, and X. Niu, "Finger-Vein Verification Using Gabor Filter and SIFT Feature Matching," in *8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, pp. 45–48, IEEE, 2012.

[39] H. Qin, L. Qin, L. Xue, X. He, C. Yu, and X. Liang, "Finger-Vein Verification Based on Multi-Features Fusion," *Sensors*, vol. 13, no. 11, pp. 15048–15067, 2013.

[40] S. Radzi, M. Khalil-Hani, and R. Bakhteri, "Finger-Vein Biometric Identification Using Convolutional Neural Network," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 24, no. 3, pp. 1863–1878, 2016.

[41] P. Y. Simard, D. Steinkraus, J. C. Platt, *et al.*, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis.," in *ICDAR*, vol. 3, pp. 958–962, 2003.

[42] B. T. Ton and R. N. J. Veldhuis, "A High Quality Finger Vascular Pattern Dataset Collected Using a Custom Designed Capturing Device," in *International Conference on Biometrics (ICB)*, pp. 1–5, Jun 2013.

[43] H. Hong, M. Lee, and K. Park, "Convolutional Neural Network-Based Finger-Vein Recognition Using NIR Image Sensors," *Sensors*, vol. 17, no. 6, pp. 1–21, 2017.

[44] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[45] H. Huang, S. Liu, H. Zheng, L. Ni, Y. Zhang, and W. Li, "DeepVein: Novel Finger Vein Verification Methods Based on Deep Convolutional Neural Networks," in *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pp. 1–8, Feb 2017.

[46] Y. Ye, L. Ni, H. Zheng, S. Liu, Y. Zhu, D. Zhang, W. Xiang, and W. Li, "FVRC2016: The 2nd Finger Vein Recognition Competition," in *2016 International Conference on Biometrics (ICB)*, pp. 1–6, IEEE, Jun 2016.

[47] C. Xie and A. Kumar, "Finger Vein Identification Using Convolutional Neural Network and Supervised Discrete Hashing," *Deep Learning for Biometrics*, pp. 109–132, 2017.

[48] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation with Noisy Labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[49] Y. Fang, Q. Wu, and W. Kang, "A Novel Finger Vein Verification System Based on Two-Stream Convolutional Network Learning," *Neurocomputing*, vol. 290, pp. 100–107, 2018.

[50] S. Zagoruyko and N. Komodakis, "Learning to Compare Image Patches via Convolutional Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353–4361, June 2015.

[51] E. Jalilian and A. Uhl, "Finger-Vein Recognition Using Deep Fully Convolutional Neural Semantic Segmentation Networks: The Impact of Training Data," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–8, IEEE, 2018.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[53] G. Lin, A. Milan, C. Shen, and I. D. Reid, "RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, p. 5, 2017.

[54] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 2481–2495, 2017.

[55] W. Kim, J. Song, and K. Park, "Multimodal Biometric Recognition Based on Convolutional Neural Network by the Fusion of Finger-Vein and Finger Shape Using Near-Infrared (NIR) Camera Sensor," *Sensors*, vol. 18, no. 7, p. 2296, 2018.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Jun.

[57] J. Wang, Z. Pan, G. Wang, M. Li, and Y. Li, "Spatial Pyramid Pooling of Selective Convolutional Features for Vein Recognition," *IEEE Access*, 2018.

[58] L. Zhang, Z. Cheng, Y. Shen, and D. Wang, "Palmprint and Palmvein Recognition Based on DCNN and A New Large-Scale Contactless Palmvein Dataset," *Symmetry*, vol. 10, no. 4, p. 78, 2018.

[59] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning," in *31st AAAI Conference on Artificial Intelligence*, vol. 4, p. 12, 2017.

[60] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Computing Surveys*, vol. 51, no. 3, pp. 65:1–65:34, 2018.

[61] M. A. Pagnutti, R. E. Ryan, G. J. Cazenavette V, M. J. Gold, R. H. Edward Leggett, and J. F. Pagnutti, "Laying the Foundation to Use Raspberry Pi 3 V2 Camera Module Imagery for Scientific and Engineering Purposes," *Journal of Electronic Imaging*, vol. 26, no. 1, pp. 1–13, 2017.

[62] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A Toolbox for Intensity-Based Medical Image Registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[63] A. Chalmers, P. Campisi, P. Shirley, and I. G. Olaizola, *High Dynamic Range Video: Concepts, Technologies and Applications*. Academic Press, 2016.

[64] J. Kuang, G. M. Johnson, and M. D. Fairchild, "iCAM06: A Refined Image Appearance Model for HDR Image Rendering," *Journal of Visual Communication and Image Representation*, vol. 18, no. 5, pp. 406–414, 2007.

[65] E. Piciucco, E. Maiorana, and P. Campisi, "Palm Vein Recognition Using a High Dynamic Range Approach," *IET Biometrics*, vol. 7, no. 5, pp. 439–446, 2018.

[66] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[67] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, p. 3, 2017.

[68] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[69] T. Contributors, "TorchVision Models." https://pytorch.org/docs/stable/torchvision/models.html, 2018.

[70] A. Canziani, A. Paszke, and E. Culurciello, "An Analysis of Deep Neural Network Models for Practical Applications," *arXiv preprint arXiv:1605.07678*, 2016.

[71] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, 2015.

Rıdvan Salih Kuzu received the bachelor's degree in Electrical & Electronics Engineering in 2010, and the master's degree in System & Control Engineering in 2017, at Boğaziçi University, İstanbul, Turkey. Since 2010, he performed different activities with roles in R&D engineering, consultancy, and management for private IT and R&D companies. Since 2017 he is a PhD student in Applied Electronics at Roma Tre University, participating in the Marie Sklodowska-Curie H2020 AMBER project. His current research interests are in biometric recognition, signal processing, pattern recognition, machine learning, and information retrieval on textual/visual data.

Emanuela Piciucco received the bachelor's degree in Electronic Engineering (cum laude) in 2013, and the master's degree in Information and Communication Technology Engineering (cum laude) in 2016, at Roma Tre University, Rome, Italy, where she is currently a PhD student in Applied Electronics. She was a Visiting Researcher at University of Salzburg, Salzburg, Austria, in 2015, in the framework of the European project ICT COST Action IC1206, and at Telefonica I+D, Barcelona, Spain, in 2017 and 2018, in the framework of the European project ENCASE. Her current research areas are biometric recognition, mainly focusing on vein pattern and EEG biometric identifiers, and physiological signal processing.

Emanuele Maiorana (IEEE SM) received the Ph.D. degree in biomedical, electromagnetism, and telecommunication engineering with European Doctorate Label from Roma Tre University, Rome, Italy, in 2009. He is currently a Research Engineer with the Section of Applied Electronics, Department of Engineering, Roma Tre University, Rome, Italy. His research interests are in the area of digital signal and image processing, with specific emphasis on biometric recognition. He is an Associate Editor of the IEEE Transactions on Information Forensics and Security. He is the recipient of the Lockheed Martin Best Paper Award for the Poster Track at the IEEE Biometric Symposium 2007, and the Honeywell Student Best Paper Award at the IEEE Biometrics: Theory, Applications and Systems conference 2008.

Patrizio Campisi (IEEE SM) received the Ph.D. degree in electrical engineering from Roma Tre University, Rome, Italy, where he is currently a Full Professor with the Section of Applied Electronics, Department of Engineering. His current research interests are in the area of biometrics and secure multimedia communications. He is a co-recipient of the IEEE ICIP06 and the IEEE BTAS 2008 Best Student Paper Award and the IEEE Biometric Symposium 2007 Best Paper Award. He was the IEEE SPS Director Student Services (2015 - 2017) and the Chair of the IEEE Technical Committee on Information Forensics and Security (2017 - 2018). He is a member of the IEEE Technical Committee on Information Assurance and Intelligent Multimedia-Mobile Communications, System, Man, and Cybernetics Society, and was a member of the IEEE Certified Biometric Program Learning System Committee. He was the General Chair of the 26th European Signal Processing Conference EUSIPCO 2018, Italy, of the 7th IEEE Workshop on Information Forensics and Security (WIFS) 2015, Italy, and of the 12th ACM Workshop on Multimedia and Security 2010, Italy. He was Technical Co-Chair of the First ACM Workshop on Information Hiding and Multimedia Security 2013, France, and of the Fourth IEEE WIFS 2012, Spain. He is the Editor of the book Security and Privacy in Biometrics (Springer, 2013). He is a Co-Editor of the books Blind Image Deconvolution: Theory and Applications (CRC press, 2007), and High Dynamic Range Video, Concepts, Technologies and Applications (Academic Press, 2016). He was an Associate Editor and a Senior Associate Editor of the IEEE Signal Processing Letters, and an Associate Editor of the IEEE Transactions on Information Forensics and Security. He is currently Editor-in-Chief of the IEEE Transactions on Information Forensics and Security.